不確かさ評価の最新動向に関する調査研究

高井絢之介*

（2024 年 1 月 31 日受理）

Investigation Report of the Latest Research for Uncertainty Evaluation

TAKAI Junnosuke

Abstract

近年機械学習を用いて，物理量の測定をより良くしようという動きがしばしば見受けられる．しかし機械学習を測定に用いる際に，計量トレーサビリティの確保や精密測定において必要不可欠である不確かさの評価をどのようにして行うかということ，より具体的には aleatoric uncertainty と epistemic uncertainty と呼ばれる不確かさの評価をどのようにするかということが新たな課題となっている．本調査報告では最新の不確かさ評価に関する研究の中で，最近発展の目覚ましい機械学習を用いて得た測定結果に対する不確かさ評価について，その課題と Gaussian Process，MC Dropout，Deep Ensemble の三つの手法を中心に課題解決が期待されているいくつかの手法，今後の研究展望についての報告を行う．

1. はじめに

何かの物理量を測定した時，その測定値は測定対象量の真の値そのものを表しているとは限らない．測定された値は，測定値の母平均に対してばらつきを持っており，またその母平均もほとんどの場合何らかの原因で真の値から離れている．そもそも真の値自体も一つの値に定まっているとは限らない．もちろん測定対象量の真の値は知ることは基本的に難しい（SI で定義される物理量のような場合を除く）が，測定値が真の値からどれだけの広がり（信頼性）を持って与えられているかを評価することはできる．この広がりのことを「不確かさ」といい，ある測定においてこの広がりを評価することを「測定の不確かさ評価」という．測定の不確かさ評価は計量トレーサビリティの確保や精密測定において必要不可欠である．

一方で機械学習は機械に学習能力を与える技術とその理論的背景を探求する研究領域である[1]．具体的には物理モデルがわからない，あるいは計算複雑性の高い測定モデルに対して，データから測定モデルを構築していくことをここでは機械学習と呼ぶ．機械学習は様々な領域で応用されているが，近年測定の分野においても機械学習を応用しようという研究が多くみられる．

しかし，機械学習を測定に用いるにあたって，その測定の不確かさ評価をどう行うかというのが一つの問題となっている．つまり機械学習を用いた測定において，測定値として得られるのは機械学習モデルからの出力であるが，この出力がどの程度の不確かさを持っているかということを定量的に評価することが難しいということである．この問題は従来の不確かさ評価法を機械学習で得られた結果に適用しても，aleatoric uncertainty と epistemic uncertainty[2),3)] と呼ばれる二つの不確かさをきちんと評価をすることができないという考えから生まれている．ここで大まかにいえば，aleatoric uncertainty は，ランダム性，つまり実験結果の変動性を表しており，追加情報を与えても低減できないような確率的成分が含まれている．また epistemic uncertainty は最適なモデルについての知識不足による不確かさを表しており，基本的なランダム現象ではなく，無知によって引き起こされるものであり，原則的に追加情報によって低減することができる．次に，aleatric uncertainty と epistemic uncertainty を従来の回帰手法で得られた結果に対する不確かさ評価法と比較して簡単に説明する．今までの回帰手法と機械学習と呼ばれている手法との大きな違いの一つは，人間が測定モデル（回帰モデル）をある程度限

* 工学計測標準研究部門 データサイエンス研究グループ

定したうえで回帰をするか，かなり弱い制限でデータから測定モデルを構築していくかという，いわゆるモデルに対する事前制限の強さだと考えられる．今までの回帰問題であれば測定モデルがデータに対して正しい挙動を示すことは仮定されていて，あてはめ自体の不確かさとデータそのものの持つ不確かさというのを計算していたが，機械学習ではデータからモデルを構築していくため，本当にモデルが正しい構造をしているのかということの評価が難しいという問題が生じる．加えてモデルの制限が弱いということはそれだけモデル構造は複雑になるはずで，あてはめがうまくいっているのかどうかという部分の定量化も一般の回帰と比べて難解であることは言うまでもない．あてはめ自体の不確かさとモデルの不確かさを合わせたものが epistemic uncertainty，データそのものが持つ不確かさ（データのばらつきなど）が aleatoric uncertainty と解釈できる．

この調査研究報告では測定に機械学習を用いた際の不確かさ評価手法について紹介する．2 章では基礎知識として機械学習に関する概要と機械学習を用いた際に評価するべき二つの不確かさ，aleatoric uncertainty と epistemic uncertainty について説明する．3 章では 2 章で説明をした不確かさ評価を行うことができると期待されている手法の中でも，主要な三つの手法を紹介し，その後他の手法についても簡単に紹介する．4 章では機械学習の手法が計量の領域において満たすべき要請と，今後解決するべき課題について議論した後，5 章で本調査研究のまとめを行う．

2. 基礎知識

ここでは機械学習，そして機械学習を測定に用いた際の不確かさ評価についての基本的な知識について紹介する．

2.1 機械学習

機械学習の代表的な機能として，分類・回帰の 2 つがあげられることが多い．学習の様式としては，正解つきの訓練データで学習する教師あり学習と，正解のついていない訓練データで学習する教師なし学習がある．一般に，分類・回帰は教師あり学習のカテゴリーに属する．

分類はデータが所属するカテゴリー（クラス）を予測する機能であり，動物の画像からその動物の種類を判別するといった例が挙げられる．機械学習を利用した分類の例としては，内視鏡，MRI，X 線画像などの医療画像から疾病を判定する診断行為の自動化がある．このような判定作業は従来医師の専門技能であったが，機械学習の活用により判定を自動化するだけではなく，熟練した医師でも発見できないような疾病の特徴を，機械学習であれば発見できるようなものも開発され始めている [4]．

回帰はデータの属性である数値を予測する機能であり，身長から体重を予測するといった問題は回帰の一例である．機械学習による回帰の実用的な例としては，機械学習を用いた株価予測 [5] のほか，製造業における工程管理 [6] にも応用がみられる．

機械学習は様々な領域で応用され始めている．例えば 2022 年に OpenAI によって公開された ChatGPT[7] は非常に高度な言語モデルとして注目を集めている．この機械学習応用の波は測定の領域においても例外ではない．例えば原子スピンの量子測定における確率過程の正確な予測 [8] や，レーザー光をとらえた CCD カメラの画像のノイズを機械学習によって低減することで測定精度を向上させる [9] などといったより良い測定をするために機械学習を測定に応用する研究が近年盛んに行われている．こういった研究では基本的に，値の予測の部分（例えば，ばねの伸びからばねにかかる力を予測する）に機械学習が利用される（図 1）．

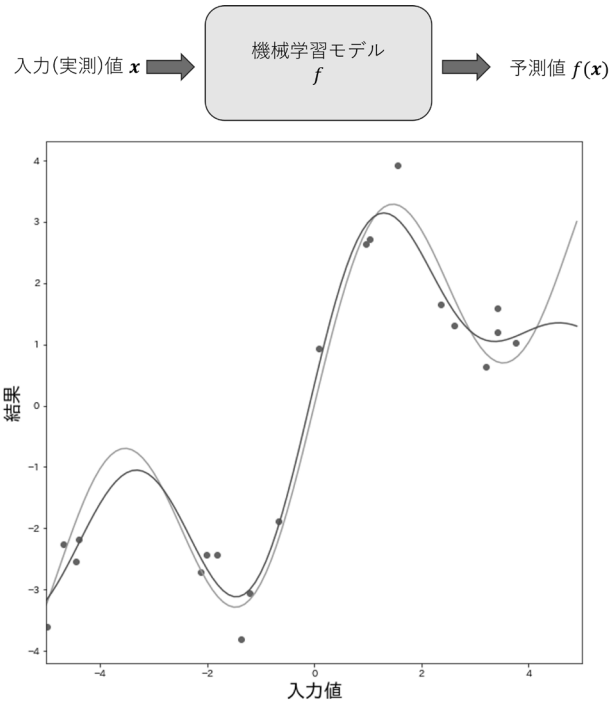

図 1 機械学習を用いた測定のイメージ図．下のグラフは実際に入力データから Gaussian Process Regression (後述) で予測値を回帰している．青点は，橙色の線 (オリジナルの関数) にノイズを付加してつくられた人工的な入力データで，赤線は予測した回帰曲線

このように測定において様々な利活用が期待される機械学習であるが，計量の領域で使用するためには不確かさ評価をする必要がある．それでは機械学習によって得られた値の予測値の不確かさの評価というのはどのようにして行えばよいのだろうか．この場合の不確かさ評価は，機械学習モデルがどの程度測定対象量を表現できているのかという不確かさを評価しなければならない．次節では機械学習を測定に用いた際の不確かさについてより詳細に議論を行う．

2.2 機械学習における不確かさ

ここでは基本的な教師あり学習の解説をした後，それを踏まえて機械学習における不確かさを見ていこう[10]．まず最初にインスタンスを一回の測定における変数の集合とし，ある特定の測定によって得られたインスタンスを query インスタンスと呼ぶことにする．教師あり学習ではまず訓練データセット D が与えられる．

$$\mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\} \subset \mathcal{X} \times \mathcal{Y}.$$

ここで X はインスタンス空間，Y はインスタンスに対応した結果の集合である（インスタンスは一回の測定における変数の集合）．ここで訓練データ（x_i, y_i）は $X \times Y$（X と Y の直積空間）上の未知の確率密度関数 P によって独立同分布（i.i.d）で生成される（全て同じ確率密度関数によって独立にデータが生成される）とする．また出力は単一であるとする．仮説空間 H（x から y への写像 $h : X \to Y$ からなる空間）と損失関数 $l : Y \times Y \to \mathbb{R}$ を仮定すると，リスク $R(h)$ は次で定義される．

$$R(h) : = \int_{\mathcal{X}\times\mathcal{Y}} l(h(\boldsymbol{x}), y) dP(\boldsymbol{x}, y).$$

またリスク $R(h)$ を最小化した仮説集合を以下のように定義する．

$$\mathcal{H}^* := \underset{h\in\mathcal{H}}{\operatorname{argmin}} R(h).$$

また実際の学習においては与えられた有限個のデータから仮説を推定しなければならないので，リスク $R(h)$ は次に定義される経験的リスク $R_{emp}(h)$ によって近似される．

$$R_{\mathrm{emp}}(h) := \frac{1}{N}\sum_{i=1}^{N} l(h(\boldsymbol{x_i}), y_i).$$

ここで $R_{emp}(h)$ を最小化した仮説集合を次のように定義する．

$$\hat{\mathcal{H}} := \underset{h\in\mathcal{H}}{\operatorname{argmin}} R_{\mathrm{emp}}(h).$$

ここで実際に予測に使う仮説は一つであるので仮説集合 $\hat{H}$ の中から，学習者が一つ選んだ仮説を経験的リスク最小仮説 $\hat{h}$ とする．この仮説は例えば学習者が経験的リスクを最小化する過程によって決まる．また今仮説空間を十分に広く，かつリスクを妥当なモデルが選択されるようにとっていると仮定し，リスクを最小化した仮説集合を一要素からなる集合とする．

$$\mathcal{H}^* = \{h^*\}.$$

この要素 h^* をリスク最小仮説とする．ここで $R_{emp}(h)$ は真のリスク $R(h)$ の推定値に過ぎないので経験的リスク最小仮説はリスク最小仮説とは基本的に一致しない．よって $\hat{h}$ が h^* を正確に推定できないことによる不確かさ（偏差）が生じる．

また我々は最終的に予測に関する不確かさ，つまり具体的なインスタンス $x_q \in X$ に対する予測 $\hat{y}_q$ に関する不確かさに興味がある．実際，個々のインスタンスに対応する推定値と不確かさを定量化することは，平均的な不確かさの定量化よりも重要な場合がある[11]．

ここで X と Y の間の依存性は決定論的ではないため，インスタンス x_q が与えられた時の y_q は Y 上の条件付き確率密度関数

$$P(y_q|\boldsymbol{x}_q) = \frac{P(y_q, \boldsymbol{x}_q)}{P(\boldsymbol{x}_q)}$$

で与えられ，y_q の値は一意に決まらず条件付確率密度関数に従って与えられる．つまり，確率密度関数 P に関する完全な情報を知りえたとしても y には不確かさが生じる．Hüllermeier らはこの不確かさを aleatoric uncertainty と呼んでいる[10]．ここで最良予測集合 F^* を次のように定義する．

$$\mathcal{F}_q^* := \underset{\hat{y}\in\mathcal{Y}}{\operatorname{argmin}} \int_{\mathcal{Y}} l(y, \hat{y}) dP(y|\boldsymbol{x}_q).$$

ここでリスク最小仮説と同様に，損失関数が妥当なモデルを選択できるような適正な関数であるとして，最良予測集合の要素は単一だと仮定する．すなわち

$$\mathcal{F}_q^* = \{f^*(\boldsymbol{x}_q)\}$$

となる．この $f^*(x_q)$ を最良予測とする．ここでリスク最小仮説 $h^*(x_q)$ と最良予測 $f^*(x_q)$ は一般的に一致しない．この間に生じる違い（二つの仮説によって出力された予測値の偏差）は，どれだけ正しいモデルを考えられているか，つまり H の選択に関する不確かさとして評価される．この不確かさは model uncertainty と呼ばれている [10]．

また学習 algorithm によって生まれた予測 $\hat{h}$ は，あくまで h^* の近似に過ぎず，その近似精度は訓練データの質と量に依存する．$\hat{h}$ と h^* の間の違いは，どれだけ近い近似ができているかという不確かさとして評価される．この不確かさは approximation uncertainty と呼ばれている [10]．そして approximation uncertainty と model uncertainty を足し合わせたものが epistemic uncertainty と呼ばれている [10]（図 2）．

aleatoric uncertainty と epistemic uncertainty を区別する一つの基準は，追加情報によってその不確かさが低減できるかどうかということである．aleatoric uncertainty はインスタンス x と結果 y の関係が確率的であり，入出力関係が非決定的であることに起因するため，追加情報によって低減不可能な不確かさである．対して，epistemic uncertainty（model uncertainty, approximation uncertainty）は完璧な予測に対する知識の欠如に起因する不確かさであり，追加情報によって低減可能である．

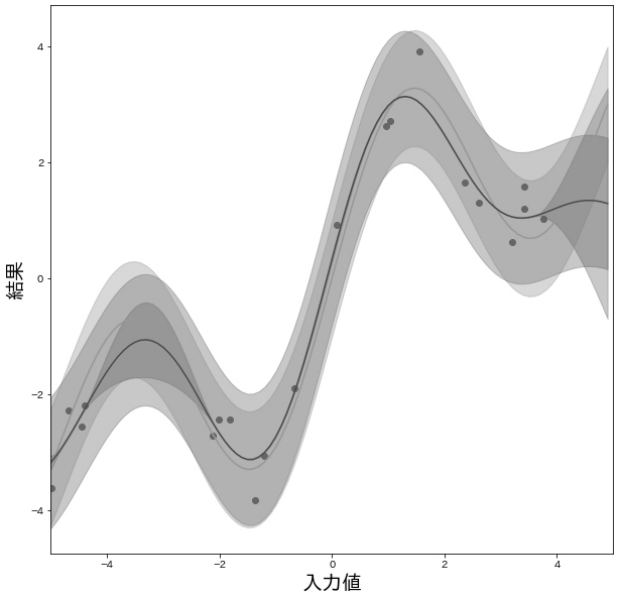

図 2　不確かさ評価を含めた機械学習回帰．図 1 と同じデータ点に対して回帰を行っている．青点は，橙色の線（オリジナルの関数）にノイズ（灰色の帯）を付加してつくられた人工的な入力データで，赤線は予測した回帰曲線，緑色の帯は aleatric uncertainty，赤色の帯は epistemic uncertainty を表している．帯の幅はすべて標準偏差の 2 倍の大きさを表している．データ点の少ない部分で epistemic uncertainty が高くなっていることがわかる．

低減可能性についてもう少し詳しく考えてみよう．学習者が特徴を増やすことによって，インスタンスの次元を増やしたならば，X を特徴の追加に対応したインスタンス空間 X' に置き換えることに相当する．この変更は不確かさに対して影響を及ぼす．例えば図 3 上のように低次元領域で二つのクラスの分布が重なっているような場合，この重なりが aleatoric uncertainty を生んでいる．しかし，図 3 下のようにデータを高次元領域に埋め込む（例えば新たな特徴量を追加する）ことで，分離可能な分布になり，aleatoric uncertainty は低減することができる．一般的には，データを高次元空間へ埋め込むことで，aleatoric uncertainty を減らすことができる一方，model のあてはめが難しくなり，より多くのデータが必要になることから epistemic uncertainty は増加するといえる．

この例からわかるように，aleatoric uncertainty と epistemic uncertainty は絶対的な概念ではなく，(X, Y, H, P) の設定に依存する．もし学習者に設定の変更を許した場合，この二つの不確かさの定量化はより困難になる．

(X, Y, H, P) を固定して考えると，学習者の知識不足というのはデータセットの大きさによって決まる．データの数 $N = |D|$ が大きくなればなるほど，学習者の無知は減っていくだろう．そして，$N \to \infty$ の極限で学習者は h^* を同定すること，つまり approximation uncertainty を完全に消すことができる．

ここで model uncertainty を捉えるためには，仮説空間 H の中の真の仮説 h についての不確かさではなく，仮説空間の候補の集合 $\mathbb{H}$ の中の正しい H についての不確かさを考えなければならないため，非常に難しいことに注意したい．モデルが正しく選択されているという仮定をすることは，モデルの誤選択のリスクを無視していることになるが，実用上はこのような仮定の下で学習を行うことが多い．

例えば Deep Neural Networks と呼ばれる機械学習モデルのような表現力の高い（H が非常に大きい）モデルでは $h^* = f^*$ または $h^* \approx f^*$ を仮定することができる．言い換えればモデルの仮定を弱くすることで，モデルの不確かさを減らしているということである．だが，モデルの仮定を弱くしたことによって approximation uncertainty が大きくなることには注意したい．

他の設定が変化する場合，例えば訓練データとテストデータの分布の間に大きな違いがある場合や，確率密度関数Pが非定常であるような場合はこれも設定の固定という仮定が成り立たなくなるので注意が必要である．

3. 不確かさ評価可能な機械学習手法

この章では先ほど見たaleatoric uncertaintyとepistemic uncertainty の二つをきちんと評価できると期待される手法についてみていく．まず最初に最近研究が進んでいる Bayes 機械学習と呼ばれる Gaussian Process Regression，Bayesian Neural Network，そして Bayes 機械学習ではないが不確かさを評価することのできる Deep Ensemble の三つについて紹介した後，今後の発展が望まれる他手法について簡単に紹介していく．

3.1 Bayes 機械学習

まず Gaussian Process Regression，Bayesian Neural Network において重要な概念である Bayes 機械学習について紹介していく．Bayes 推定では，確率的な予測からなる仮説空間 H を考え，仮説 h はインスタンス x が与えられた時の結果 y の確率密度関数 $p_h(y|x) = p(y|x, h)$ を出力する．Bayes の approach では，事前分布 $p(\cdot)$ が用意され，この事前分布をモデルの事後分布 $p(h|D)$ に

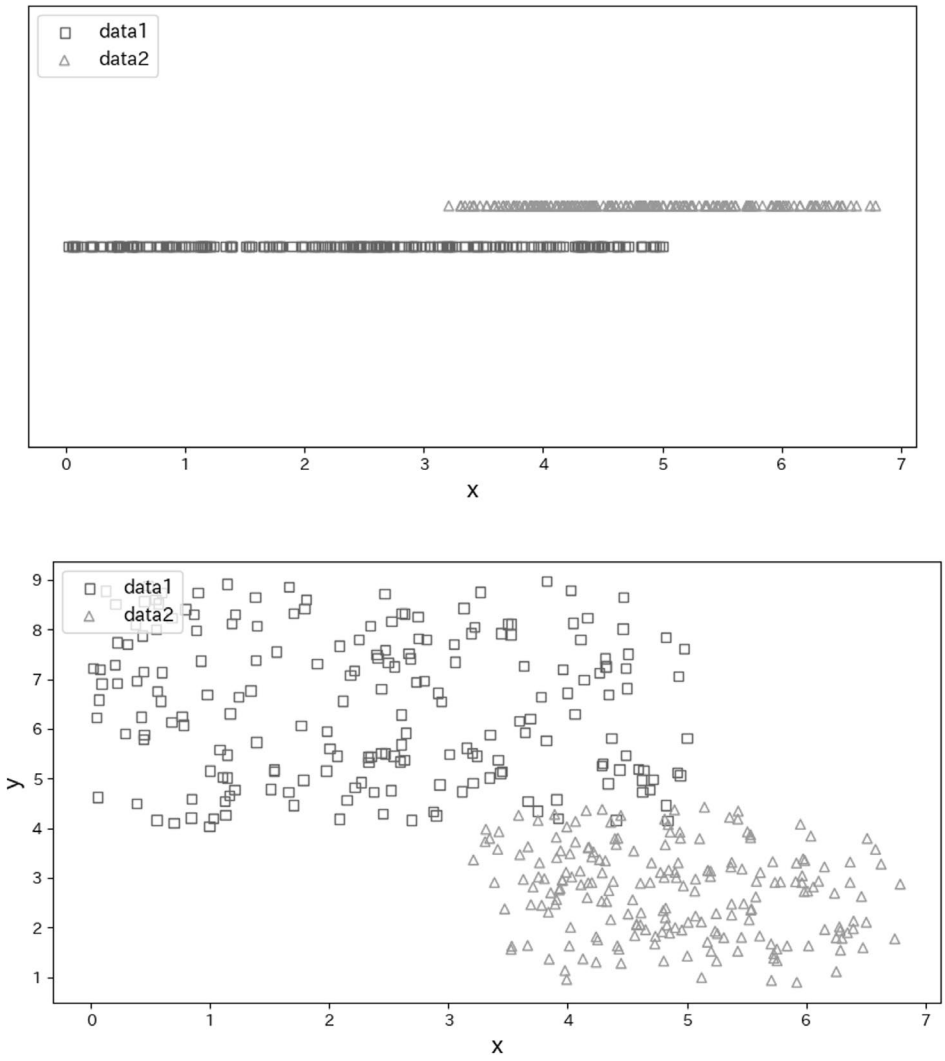

図 3 設定を変化させたときに aleatoric uncertainty と epistemic uncertainty が変化することの例．上図は二種類のデータを一次元の特徴で，下図は二次元の特徴であらわしている．

置き換える操作を行う．

$$p(h|\mathcal{D}) = \frac{p(h)\cdot p(\mathcal{D}|h)}{p(\mathcal{D})} \propto p(h)\cdot p(\mathcal{D}|h).$$

ここで $p(\mathcal{D}|h)$ は h が与えられたときのデータの確率であり，モデル h の尤度は $p(\mathcal{D}|h)$ に比例する．直感的には，$p(\cdot|\mathcal{D})$ は学習者の知識の状態であり，epistemic uncertainty を捉えている．この分布が尖っているほど，つまり H の小さい領域に確率が集中しているほど，学習者にとっての不確かさは小さくなる．H 上の事後分布は，インスタンスごとの局所的な情報ではなく，インスタンス空間全体で平均化されたグローバルな情報を表す．特定の query インスタンス $\boldsymbol{x}_q$ に対する予測 y_q の不確かさの表現は，以下の予測事後分布によって与えられる．

$$p(y|\boldsymbol{x}_q) = \int_{\mathcal{H}} p(y|\boldsymbol{x}_q, h)dP(h|\mathcal{D}).$$

Bayes 推定では，最終的な予測はモデル平均によって生成される．つまり，結果の確率は，各仮説 h の確率の重みでとった，H 内のすべての仮説による加重平均となる．

予測事後分布において，aleatoric uncertainty と epistemic uncertainty は区別されなくなり，epistemic uncertainty が平均化される．これを説明するためにコイントスの例を考えてみる．コインが表になる確率を α，その時のコインのモデルを h_α として，仮説空間を $H := \{h_\alpha | 0 \leq \alpha \leq 1\}$ とする．この時，事後分布 P が一様分布の場合（全てのモデルが同じ確率，つまり完全な無知の場合）と事後分布 P' が，$h_{1/2}$ が確率 1 になるような分布の場合 P'（コインが完全に fair であることが確実な場合）で表と裏の確率は両方とも 1/2 になる．

$$p(y) = \int_{\mathcal{H}} \alpha dP = \frac{1}{2} = \int_{\mathcal{H}} \alpha dP'.$$

より一般的に，$Y := \{-1, +1\}$ の二値分類を考える．$p_h(+1|\boldsymbol{x}_q)$ を仮説 h によって +1 に分類される確率として，Y 上の事後分布を予測する代わりに，未知の確率 $q := p(+1|\boldsymbol{x}_q)$ の予測事後分布を導出する．

$$p(q|\boldsymbol{x}_q) = \int_{\mathcal{H}} [[p(+1|\boldsymbol{x}_q, h) = q]]dP(h|\mathcal{D}).$$

ここで，$[[a = b]]$ は $a = b$ の時 1 を，$a \neq b$ の時 0 をとるような定義関数である．上式は二次の確率（確率の確率）であり，aleatoric uncertainty と epistemic uncertainty を両方含んでいる．直感的には epistemic uncertainty はこの分布の変動性に反映される，しかし，この概念をどのように定量化するのかというのは一つの課題である．

次のセクションでは二つの主要な Bayes 機械学習の手法について紹介する．

3.1.1 Gaussian Process Regression

最尤推定と同様に，パラメータ $\theta \in \Theta \subseteq \mathbb{R}^d$ によってそれぞれの仮説 $h = h_\theta \in H$ が一意に決まるという仮定の下で，Bayes 推定の仮説空間はパラメタライズされる．したがって，事後分布の計算は，最良なパラメータに関する確率を更新していく作業であり，これは多変量の確率変数として扱われる．

$$p(\theta|\mathcal{D}) \propto p(\theta)\cdot p(\mathcal{D}|\theta).$$

Gaussian Process Regression[12] は多変量のランダム変数に関する Bayes のアプローチを，（無限次元の）関数に関する推論に一般化している．したがってそれらはランダムベクトルだけでなく，ランダム関数に関する分布としても考えることができる．

より具体的には，有限の要素の集合 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in X$ に対して，対応するランダム変数の集合は $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m)$ は以下のように平均値 $m(\boldsymbol{x})$，共分散 $k(\boldsymbol{x}, \boldsymbol{x}')$ の多変量正規分布で表される．

$$\begin{bmatrix} f(\boldsymbol{x}_1) \\ f(\boldsymbol{x}_2) \\ \vdots \\ f(\boldsymbol{x}_m) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\boldsymbol{x}_1) \\ m(\boldsymbol{x}_2) \\ \vdots \\ m(\boldsymbol{x}_m) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_m) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_m, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_m, \boldsymbol{x}_m) \end{bmatrix} \right).$$

Gaussian Process Regression を使用することで，行列の計算で正しい Bayes 推定を行うことができる．関数 $h \in H$ についての事前分布を，測定値 $y_i = f(\boldsymbol{x}_i) + E_i$ で構成されるデータ $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ による事後分布へと変換していく．同様に新しい query インスタンス $\boldsymbol{x}_q \in X$ に対する結果 $y \in Y$ の予測事後分布が得られる．一般的な場合では，この計算は非常に困難だが，ガウスノイズを持つ回帰の場合は簡単に計算することができる．多変量正規分布の平均値 $m(\boldsymbol{x})$ を 0 とすると，予測事後

分布は以下の平均 μ，分散 σ^2 のガウス分布によって与えられる．

$$\begin{aligned}\mu &= K(\boldsymbol{x}_q, X)(K(X,X)+\sigma_\epsilon^2 I)^{-1}\boldsymbol{y},\\ \sigma^2 &= K(\boldsymbol{x}_q, \boldsymbol{x}_q)+\sigma_\epsilon^2\\ &\quad - K(\boldsymbol{x}_q, X)(K(X,X)+\sigma_\epsilon^2 I)^{-1}K(X, \boldsymbol{x}_q).\end{aligned}$$

ここで X は訓練データからなる $N \times d$ 行列，$K(X, X)$ は要素 $(K(X, X))_{i,j} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ であるようなカーネル行列，$\boldsymbol{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ は測定された結果のベクトル，σ_ϵ^2 は測定のノイズによる分散である．

回帰の場合，query インスタンス $\boldsymbol{x}_q$ の予測事後分布の分散 σ^2 は意味のある不確かさの指標である．測定による分散 σ_ϵ^2 は，aleatoric uncertainty に対応しているため，その差は epistemic uncertainty と考えることができる．

3.1.2 Bayesian Neural Network

標準的な Neural Network（NN）は，分類の場合は確率密度関数がクラス集合 Y 上で出力され，回帰の場合は期待値とみなされる点予測 $h(\boldsymbol{x}) \in \mathbb{R}$ が出力される．NN の訓練は基本的に最尤推定に基づいて行われる．そのため，aleatoric uncertainty は捉えるが，epistemic uncertainty は捉えられない．

NN の文脈では，モデルの高い自由度のために model uncertainty は無視できると考えられるため，epistemic uncertainty は一般的にモデルパラメータ（重み w に対応するもの）に関する不確かさとして理解されていて，これは approximation uncertainty と基本的には同等である．この epistemic uncertainty を捉えるために，Bayesian Neural Network（BNN）が提案された[13)~15)]．BNN では，各重みは実数ではなく確率密度関数で表現され，学習は Bayes 推定に帰結し，事後分布 $p(w|D)$ を計算する．query インスタンス $\boldsymbol{x}_q$ が与えられたときの出力の予測分布は次のようになる．

$$p(y|\boldsymbol{x}_q, \mathcal{D}) = \int p(y|\boldsymbol{x}_q, \boldsymbol{w}) p(\boldsymbol{w}|\mathcal{D}) d\boldsymbol{w}.$$

重みに関する事後分布を解析的に計算することは通常はできないので，変分的な手法を用いた近似的な手法が利用されている[16),17)]．これらの手法は，重みの近似事後分布 q_θ と真の事後分布 $p(w|D)$ の間の Kullback-Leibler Divergence $KL(q_\theta \| p(w|D))$ を最小化する q_θ を求める．Gal と Ghahramani の提案した Monte Carlo（MC）Dropout は変分的な近似の一つの重要な例である[18)]．MC Dropout は Dropout と呼ばれる手法を用いて，BNN の近似を行うことのできる手法である．Dropout はランダムに選ばれたネットワークのノードのみで学習を行うことで，過学習を防ぐ正則化手法のうちの一つである．本来学習時のみに Dropout を作用させ，予測をする際はすべてのネットワークパラメータを用いるが，MC Dropout では学習時に Dropout を用いるだけでなく予測時にもランダムにノードを落とす（図 4）．この出力を落とすノードを変えて何回も繰り返すことで，出力される値は一つの値ではなく繰り返し回数分の値の集合になる．この集合が近似的に BNN から出力される予測事後分布の標本集合としてみなされるというのが MC Dropout の理論である．つまりこの出力の平均と標準偏差を計算すれば，BNN から出力される予測事後分布の期待値と標準偏差を近似的に推定できるということである．ここで予測時だけでなく学習時にも Dropout を用いて学習を行わなければならないことに注意したい．

BNN によって出力される予測事後分布の不確かさは，aleatoric uncertainty と epistemic uncertainty の両方を含む．前者は定数ではなく $\boldsymbol{x} \in X$ の関数である可能性があることに注意したい．最近 $\boldsymbol{x} \in X$ に依存するような aleatoric uncertainty を学習する方法も研究されている．

Depeweg[19)] は明示的に aleatoric uncertainty と epistemic uncertainty を分離しようとしている．つまり予測事後分布のエントロピーを使用して合計の不確かさと aleatoric uncertainty を評価し，epistemic uncertainty をその差として取得する．より具体的には，離散的な Y の場合は合計の不確かさは以下のようになる．

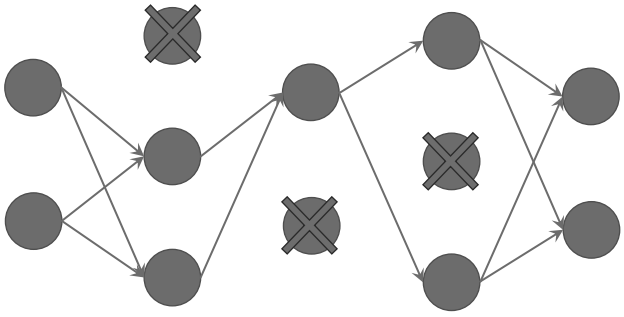

図 4 MC Dropout のイメージ図．複数のノードと層からなる NN だが，そのノードを学習時と予測時にランダムに落とすことによって出力される値をばらつかせる．その平均と標準偏差が Bayes による予測事後分布の期待値と分散に近似的に一致する．

$$H\left[p(y|\boldsymbol{x})\right] = -\sum_{y\in\mathcal{Y}} p(y|\boldsymbol{x})\log_2 p(y|\boldsymbol{x}).$$

この不確かさには重みに関する不確かさ（epistemic uncertainty）が含まれているが，$\boldsymbol{p}(y|\boldsymbol{w},\boldsymbol{x})$のエントロピーを，重みに関して期待値をとれば epistemic 不確かさを除去した不確かさを得られる．

$$E_{p(\boldsymbol{w}|\mathcal{D})}H\left[p(y|\boldsymbol{w},\boldsymbol{x})\right] = -\int p(\boldsymbol{w}|\mathcal{D})\left(\sum_{y\in\mathcal{Y}} p(y|\boldsymbol{w},\boldsymbol{x})\log_2 p(y|\boldsymbol{w},\boldsymbol{x})\right)d\boldsymbol{w}$$

は aleatoric uncertainty の尺度である．最終的に epistemic uncertainty は以下のようにして計算される．

$$u_{\mathrm{e}}(\boldsymbol{x}) = H\left[p(y|\boldsymbol{x})\right] - E_{p(\boldsymbol{w}|\mathcal{D})}H\left[p(y|\boldsymbol{w},\boldsymbol{x})\right].$$

この量は y と w の相互情報量 $I(y,w)$ に等しい．直感的には，epistemic uncertainty は真の結果 y の知識を通じてモデルパラメータ w に関する情報の量を捉える．もちろんこのエントロピーの尺度は直接的に不確かさの尺度（確率密度関数の分散）には変換できないため取扱いには注意したい．

3.2 Deep Ensemble

ensemble 学習は教師あり学習の一つでいくつかの学習機（モデル）を組み合わせることによって，より高い精度と汎用性を持ったモデルを作る学習方法である．Deep Ensemble では測定対象量の分布の平均と分散を出力するような Neural Network を複数個作り，それらの NN モデルを組み合わせて予測を行う．この時の分布の分散による不確かさが aleatoric uncertainty に相当し，モデル間の分散による不確かさが epistemic uncertainty に相当する．

より具体的に見ていこう．まず想定として，N 個の i.i.d なデータセット $\boldsymbol{D} = \{x_n, y_n\}_{n=1}^N$ を用意する．（$x \in \mathbb{R}^d$, $y \in \mathbb{R}$）次に x_p を入力して，予測値 y_p の分布の平均 $\mu_{\boldsymbol{\theta}}(x_p)$ と分散 $\sigma^2_{\boldsymbol{\theta}}(x_p)$ を出力するような NN を作り学習させる（$\boldsymbol{\theta}$ は network のパラメータ）．ここで学習に使うデータはデータセット $\boldsymbol{D}$ の中から重複を許して N 個のデータをサンプリングして，そのデータセットをモデルの訓練に使用する（これを Bootstrap 法という）．ここで $\mu_{\boldsymbol{\theta}}(x)$ と $\sigma^2_{\boldsymbol{\theta}}(x)$ はそれぞれ結果の分布の平均と分散に対応しているため，結果の分布を Gauss 分布と仮定すると，負の対数尤度は以下のようになる．

$$\begin{aligned}&-\log p_{\boldsymbol{\theta}}(y_n|x_n)\\ &= \frac{1}{2}\log\sigma^2_{\boldsymbol{\theta}}(\boldsymbol{x}) + \frac{(y-\mu_{\boldsymbol{\theta}}(\boldsymbol{x}))^2}{2\sigma^2_{\boldsymbol{\theta}}(\boldsymbol{x})} + const.\end{aligned}$$

この負の対数尤度を最小化するように $\boldsymbol{\theta}$ を学習していく．このまま最尤推定を行うと over fitting しやすいので，正則化項を加える，または何かしらの事前分布を作って MAP（maximum a posteriori）推定（事後分布の確率密度を最大にするような推定方法）を行うとより良い学習が行える．また，Gauss 分布ではない分布を仮定する場合は，特定の分布に対して尤度を計算する．

このような学習方法を M 回繰り返し，M 個の NN を作成する．これらの NN はすべて異なるデータセットで学習されていると期待されるためその構造も異なっており，もちろん出力も異なる出力となる．最終的な出力として全ての NN からの出力を平均したものを採用する．分布の形は混合ガウス分布となるが，その平均と分散は近似して以下のようになる [20]．

$$\mu_{\mathrm{e}}(\boldsymbol{x}) = \frac{1}{M}\sum_m^M \mu_{\boldsymbol{\theta}_m}(\boldsymbol{x}),$$

$$\sigma^2_{\mathrm{e}}(\boldsymbol{x}) = \frac{1}{M}\sum_m^M (\sigma^2_{\boldsymbol{\theta}_m}(\boldsymbol{x}) + \mu^2_{\boldsymbol{\theta}_m}(\boldsymbol{x})) - \mu^2_{\mathrm{e}}(\boldsymbol{x}).$$

ここで，$\mu_{\boldsymbol{\theta}m}(\boldsymbol{x})$ は m 個目の NN の $\mu_{\boldsymbol{\theta}}(\boldsymbol{x})$，$\sigma^2_{\boldsymbol{\theta}m}(\boldsymbol{x})$ は m 個目の NN の $\sigma^2_{\boldsymbol{\theta}}(\boldsymbol{x})$ である．また aleatoric uncertainty は $\frac{1}{M}\sum_m^M \sigma_{\boldsymbol{\theta}m}(\boldsymbol{x})$ で，epistemic uncertainty は $\frac{1}{M}\sum_m^M \mu^2_{\boldsymbol{\theta}m}(\boldsymbol{x}) - \mu^2_e(\boldsymbol{x})$ とする．

3.3 その他の手法

ここでは，その他の手法として，頻度統計に基づくパラメータ推定の考えに基づく方法と，集合論に基づく方法について紹介する．

3.3.1 最尤推定とフィッシャー情報量

尤度の概念は一般的な統計推論の中でも重要な要素であり，最尤推定は機械学習でも普遍的な原則である．実際多くの頻度論的機械学習のアルゴリズムは尤度の最大化によって成り立っている．また，尤度に基づく推定と Bayes 推定の間には密接な関係があり，無情報事前分布を利用した Bayes 推定は最尤推定と同等であることが知られている．

ここでは，$\boldsymbol{\theta} \in \Theta$ によってパラメータ化された確率密度関数 $\boldsymbol{P}_{\boldsymbol{\theta}}$ によってデータが生成されていると考えよう．ここで $p_{\boldsymbol{\theta}}(\cdot)$ は $\boldsymbol{P}_{\boldsymbol{\theta}}$ の確率密度関数とする．回帰の目的の一つはパラメータ $\boldsymbol{\theta}$ を推定すること，つまり，測定

データ $D = X_1, \ldots, X_N$ に基づいて確率密度関数 P_θ を同定することである．最尤推定はこの問題に取り組むための一般的な方法である．具体的には，尤度関数，あるいは対数尤度関数を最大化することによって $\boldsymbol{\theta}$ を推定する． $X_1, \ldots, X_N$ が独立である，つまり D が $(P_\theta)^N$ に従って分布しているとき，対数尤度関数は次のように表せる．

$$l_N(\boldsymbol{\theta}) := \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(X_n).$$

最尤推定値 $\hat{\boldsymbol{\theta}}$ の分布は漸近的に正規分布に近づくことが知られている．具体的には $\sqrt{N}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ が平均 0 で共分散行列が $I_N^{-1}(\boldsymbol{\theta})$ であるような正規分布に収束する．ここで，

$$I_N(\boldsymbol{\theta}) = -\left[E_{\boldsymbol{\theta}}\left(\frac{\partial^2 l_N}{\partial\theta_i \partial\theta_j}\right)\right]_{1\geq i,j\geq d}$$

はフィッシャー情報行列（$\boldsymbol{\theta}$ における対数尤度の負の Hessian の期待値）である．（d は $\boldsymbol{\theta}$ の次元）

フィッシャー情報行列を使用することで，推定値 $\boldsymbol{\theta}$ の近似的な信頼区間を構築できる．この区間が大きいほど，真のモデルに対する不確かさが大きくなる．信頼区間のサイズは最大の尤度関数がその最大値の周りでどれだけ広がっているかと直接対応している．

機械学習の文脈では，パラメータ $\boldsymbol{\theta}$ が仮説 $h = h_\theta$ を表す場合，信頼区間は h^* に関する epistemic uncertainty，具体的には approximation uncertainty とみなすことができる．

3.3.2 仮説集合の特性に着目した評価法

epistemic uncertainty と aleatoric uncertainty の評価を行う手法として，仮説集合の特性に着目した評価方法がある．

例えば，version space 学習という評価方法では対象とする仮説空間の中で，実際に得られた訓練データを説明することが可能な仮説がとりうる空間のサイズに着目して epistemic uncertainty の評価を行う．一方で，この手法ではランダムな誤差がないことを想定している．というのは，ランダム誤差がある場合には，どれほどの低い可能性でもある仮説を棄却することが難しく，訓練データによって仮説を限定することができなくなるためである．このため，原則的にはこの手法では aleatoric uncertainty がない仮説及びデータのみを取り扱う．

この version space 学習のランダム成分を取り扱えないという弱点を補う方法として，reliable classification という方法がある．この手法で行われるステップは以下の２つである．

1. 仮説に基づいて，query インスタンス x_q を与え，各候補結果 $y \in Y$ について妥当性の度合いを導く．
2. 妥当性の度合いから aleatoric 不確かさと epistemic 不確かさを導く．

この際「度合い」の評価には正規化尤度と呼ばれるものが使用される．確率がある仮説に基づいてデータがどの程度表れやすいかを表す指標であるのに対して，尤度とはデータに基づいて仮説がどの程度妥当かを示す指標になるものである．

他にも，仮説検定という頻度統計的の中心的アプローチに基づく conformal prediction という手法が提案されている．これは version space ではその仮説によって訓練データが説明できるか，できないかという単純に二値的な判断を行うのに対して，「有意水準」として与えられる一定の確率で説明できるかできないかという判断を行うものである．これらの手法の詳細は別途付録において説明する．

これらの手法で，aleatoric uncertainty と epistemic uncertainty とされるものは，必ずしも標準偏差という形式で評価されているわけではないことには注意したい．これらの手法を回帰分析に適用し，「測定の不確かさ」評価を行うにはさらなる統計学的な議論が必要である．

4. 今後の課題

計量における機械学習の不確かさ評価を研究するにあたっての課題を，NPL[21]が七つの要請としてまとめている．その中でも著者が研究を進めていくうえで特に重要だと考えている三つの課題を選んで説明する．

一つ目の課題は入力量の不確かさを考慮すべきだということである．GUM では不確かさは入力の不確かさを伝播させて評価していたことを思い出してほしい．現在の機械学習における多くの手法は，入力量における不確かさを明示的にモデル化をしていない．線形回帰のような単純なモデルであれば解析は容易であるが，機械学習ではモデルそのものをデータから学習させるので状況はより複雑になる．もし何か固定されたモデルで入力の不確かさを評価した場合は，そのモデルのパラメータの不確かさを無視していることになる．つまりこの二つの不確かさを組み合わせて評価できる仕組みが必要である．また文献[21]によると最新の研究では入力量の Monte-Carlo sampling とパラメータの不確かさを組み合わせた

ものなどが存在する．また Deming 回帰を用いた手法なども提案されている．

二つ目の課題はスケーラビリティがあるべきだということである．Gaussian Process Regression は data driven なアプローチとして計測業界では一般的であるが，最近の不確かさ評価では深層学習に注目が集まっている．理由は二つ考えられて，一つ目は表現力が高く，よりバイアスが抑えられると期待できる部分である．そして最も重要なのが二つ目の理由で，大規模なデータに対してのスケーラビリティがあるということである．Gaussian Process Regression ではデータ数を n とした際に，$n \times n$ の行列の逆行列を計算しなければならず，これは $O(n^3)$ の計算量オーダーとなるため大規模なデータセットに対しては計算が難しい．これを解決するために Deep Gaussian Process Regression などの組み合わせた手法の研究も行われている．

三つ目の課題は原理原則を守った不確かさ評価をするべきだということである．最近の研究は大規模な問題に対してスケーラビリティを求める方向に集中しているが，これは原理的なアプローチをある程度犠牲にすることにつながる．原理的な不確かさ評価に関する問題は三つあげられる．

1. 不確かさの概念によって不確かさ評価の原理が異なるということ．最近の研究は頻度主義的なアプローチが多い（アンサンブル学習など）
2. 現在の多くのアプローチには多くの仮定を置くことによってスケーラビリティなどを実現しているということ．例えば深層学習にベイズ推論を適用する場合，事後分布や事前分布に対して計算の単純化を目的とした仮定をとることが多く，これが非現実的であるような例も少なくない．
3. スケーラビリティは近似を行うことで達成されることが多い．例えば MC Dropout もそのうちの一つである．

今後はこれら三つの課題を解決できるような機械学習モデルを開発していくことを目標にしたい．

5. まとめ

本報告では不確かさ評価に関する最新の研究の中で，機械学習を用いた測定の測定結果に対する不確かさ評価についてまとめた．測定に機械学習を用いた際には aleatoric uncertainty と epistemic uncer- tainty と呼ばれる二つの不確かさを評価する必要があり，この二つの不確かさの評価に適していると考えられるいくつかの手法を紹介した．これらの手法にはいまだ解決するべき課題も多く，実際に不確かさ評価を行うハードルは依然高いように見受けられる．提示した課題を解決し，測定における機械学習の応用を発展させるべく研究を進めていきたい．

謝辞

本調査研究報告書の作成において，工学計測標準研究部門データサイエンス研究グループの田中秀幸氏，城野克広氏には調査の議論から原稿の構成に至るまで辛抱強くご指導いただきました．ここに感謝申し上げます．また，工学計測標準研究部門データサイエンスグループの渡邊宏氏，松岡聡氏，岡本隼一氏には調査研究の発表練習や報告書の修正において非常に貴重なご意見をいただきました．厚く御礼申し上げます．

付録 A Version Space 学習

Version Space 学習[22]では入出力関係 $f^* : X \to Y$ が決定論的だという仮定を置く．つまり，

$$p(y|\boldsymbol{x}_q) = 1 \text{ if } y = f^*(\boldsymbol{x}_q)$$

となる．また訓練データはノイズフリーであるとする．よって分類器は $h(x) \in \{0, 1\}$ のように確定的な分類を行う．最後に，$f^* \in H$，つまり $h^* = f^*$(model uncertainty が 0）を仮定する．

この仮定の下で，訓練データ D に適合しない仮説 $h \in H$ は H から削除される．この結果残った仮説からなる集合 $V \in H$ は version space と呼ばれ，以下のように定義される．

$$\mathcal{V} = \mathcal{V}(\mathcal{H}, \mathcal{D}) := \{h \in \mathcal{H} | h(\boldsymbol{x}_i) = y_i \text{ for } i = 1, \ldots, N\}.$$

明らかにデータセットが拡大すれば，version space は縮小する．つまり，$V(H, D') \subseteq V(H, D)$ for $D \subseteq D'$ である．

query インスタンス x_q に対応する予測 $\hat{y}_q$ は，version space のすべての仮説 $h \in V$ についての出力となる．形式的に結果 $y \in Y$ に対する妥当性について以下のように書く．

$$\pi(y) := \max_{h\in\mathcal{H}} \min([[h \in \mathcal{V}]], [[h(\boldsymbol{x}_q) = y]]).$$

$\pi(y)=1$ は $h(\boldsymbol{x}_q)=y$ を満たす仮説 $h \in V$ が存在することを示している．よって，version space learning による予測は以下のように部分集合で与えられる．

$$Y = Y(\boldsymbol{x}_q) := \{h(\boldsymbol{x}_q)|h \in \mathcal{V}\} = \{y|\pi(y) = 1\} \subseteq \mathcal{Y}.$$

出力 $\hat{y}_q$ についてはデータ D だけでなく，仮説空間 H にも依存することに注目する．一般的に，aleatoric uncertainty と epistemic uncertainty の両方は事前知識とデータの相互作用によって決まる．学習過程が始まる際に持つ事前知識が強ければ強いほど，不確かさを減らすのに必要なデータ数は少なくなる．極端に言えば，あらかじめ真のモデルを知っていればデータ自体が不要となる．このことからもわかるように，モデルの仮定が制約的であればあるほど不確かさは小さくなる．

不確かさに関する議論に戻ると，Version Space 学習では aleatoric uncertainty は仮定していないので epistemic uncertainty が唯一の不確かさである．モデルの不確かさ（approximation uncertainty）は version space のサイズと対応しており，データセットの大きさが増加すると減少する．同様に予測の不確かさは結果の候補の集合 $Y(\boldsymbol{x}_q)$に依存している．

付録 B Reliable Classification

正規化尤度と呼ばれる概念を活用して，aleatoric uncertainty と epistemic uncertainty の区別を行うことができる．このアプローチは集合論的な推定と分布的な推定を組み合わせており，version space 学習と Bayes 推定の中間に位置するものといえる．

基本的なアイデアを説明するために，二値分類を考える．基本的なステップは以下の二つである．

・最初に，query インスタンス $\boldsymbol{x}_q$ が与えられた場合，各候補結果 $y \in Y$ について妥当性の度合いが導く．

・妥当性の度合いから aleatoric uncertainty と epistemic uncertainty の度合いを導く．

version space 学習では，仮説は可能か不可能か，結果は仮説と矛盾するかしないかの二択によって決まっていた．Senge は仮説 $h \in H$ の妥当性を $\pi_H(h) \in [0, 1]$，結果 y を仮説 $h(x)=p(y|x) \in [0, 1]$ とすることで version space 学習を一般化した [23)]．

より具体的には，正規化尤度 $\pi_H(h)$によって仮説の妥当性を表す．

$$\pi_{\mathcal{H}}(h) := \frac{L(h)}{\sup_{h'\in\mathcal{H}} L(h')} = \frac{L(h)}{L(h^{ml})}.$$

ここで，$L(h)$はデータ D における仮説 h の尤度，$h^{ml} \in H$ は最尤推定量である．したがって，妥当性は尤度に比例しており，最尤推定量の妥当性は 1 となる．

次に H 上の不確かさを query インスタンス $\boldsymbol{x}_q$ における予測についての不確かさに変換する．{+1, − 1} の二値分類において，query インスタンス $\boldsymbol{x}_q$ が与えられたときのクラス +1 の妥当性は以下のようになる．

$$\pi(+1|\boldsymbol{x}_q) := \sup_{h\in\mathcal{H}} \min(\pi_{\mathcal{H}}(h), \pi(+1|h, \boldsymbol{x}_q)).$$

ここで，$\pi(+1|h, \boldsymbol{x}_q)$は仮説 h における +1 である確率であり，以下のようになる．

$$\pi(+1|h, \boldsymbol{x}_q) := \max(2h(\boldsymbol{x}_q) - 1, 0).$$

つまり $h \le 1/2$ の場合に 0 になり，それ以降は線形に増加する．− 1 の場合も同様に

$$\pi(-1|\boldsymbol{x}_q) := \sup_{h\in\mathcal{H}} \min(\pi_{\mathcal{H}}(h), \pi(-1|h, \boldsymbol{x}_q)),$$
$$\pi(-1|h, \boldsymbol{x}_q) := \max(1 - 2h(\boldsymbol{x}_q), 0)$$

となる．

計算した妥当性 $\pi(+1) = \pi(+1|\boldsymbol{x}_q)$，$\pi(-1) = \pi(-1|\boldsymbol{x}_q)$が高いクラスを予測していく．興味深いのは，この二つの妥当性から epistemic uncertainty u_e と aleatoric uncertainty u_a が以下のように計算されるということである．

$$u_{\mathrm{e}} := \min(\pi(+1), \pi(-1)),$$
$$u_{\mathrm{a}} := 1 - \max(\pi(+1), \pi(-1)).$$

u_e は二つの妥当性が両方とも高い度合いであり，u_a は両方とも高くはない度合いであると考えられる．この二つの不確かさは $u_a + u_e \le 1$ を満たすため，合計の不確かさは 1 を超えることはない．厳密には，$\pi(y)$は妥当性の上限として解釈するべきであり，サンプルサイズの増加とともに減少していく．したがって不確かさについても同様に考えるべきである．例えば，最初にまったくデータが観測されていない場合，両方の結果は完全に確からしいため，$u_e = 1, u_a = 0$ である．よって，u_a は真の aleatoric uncertainty の下限としてみるべきである．具体的な例として公平なコイントスを考えると，時間とともに u_a は 1 に近づき，u_e は 0 に近づいていく．

より一般的に次の特殊なケースを考える．

・epistemic uncertainty が $1 : u_e = 1$ は最も高い尤度を持つ少なくとも二つの完全に確からしい仮説がある必要がある．この状況は，鋭い尤度を持たない小さいサンプルサイズの場合に発生する可能性がある．

・epistemic uncertainty がない場合：$u_e = 0$ は $\pi(+1) = 0$ か $\pi(-1) = 0$ のどちらかの場合に成り立つ．これは全ての仮説に対して，$h(x_q) \leq 1/2$ か，$h(x_q) \geq 1/2$ ということである．つまり，どちらのクラスが予測されるかモデルに関わらず決定している状況である．

・aleatoric uncertainty が $1 : u_a = 1$ は全ての仮説が両方のクラスに確率 1/2 を割り当てている状況である．

・aleatoric uncertainty がない場合：$u_a = 0$ は最も尤度の高い仮説が片方のクラスに確率 1 を割り当てているような状況である．

アルゴリズム的な側面に深く言及はしないが，これらの計算はかなり複雑になることがある．実際，最大値の計算は最適化問題を解くことになり，その複雑性は仮説空間 H に大きく依存する．

付録 C Conformal Prediction

conformal prediction[24] は頻度統計，より具体的には仮説検定に基づいている．具体的には，訓練データと query インスタンス X_{N+1} が与えられたとき，対応する $y_{N+1} = y$ について，全ての結果 $y \in Y$ の仮説検定を行う．仮説検定によって所定の信頼度で棄却された結果は除外され，仮説が棄却できない結果は prediciton set または prediction region $Y^\epsilon \subset Y$ を形成する．点予測 $\hat{y}_{N+1} \in Y$ の代わりに，確率 $1 - \epsilon$ で真の結果 y_{N+1} を含んでいる集合値予測 $Y^\epsilon \in Y$ を用いることが conformal prediction の基本的な考えである．ここで，$\epsilon \in (0, 1)$ は事前に決めた有意水準である．分類の場合は，Y^ϵ は結果のクラス $Y = y_1, \ldots, y_k$ の集合になり，回帰の場合は区間として表現される．

検定の手順を見ていこう．スコア $\alpha = f(x, y)$ を割り当てる関数，$f : X \times Y \to \mathbb{R}$ を考える．このスコアは (x, y) の奇妙さの尺度として解釈でき，スコアが高いほど，データ点 (x, y) は期待される結果から逸脱していることを意味する．この関数をデータのシーケンスに適用し，$y = y_{N+1}$ を選択するとスコアのシーケンスが生成される．

$$\alpha_1, \alpha_2, \ldots, \alpha_N, \alpha_{N+1}.$$

$\alpha_{\tau(1)} \leq \cdots \leq \alpha_{\tau(N+1)}$ となるような置換 τ によって，スコアを並び変える．y_{N+1} の選択が真のデータ生成プロセスに従っており，この過程が交換可能性（この仮定は独立性よりも弱く，ただ単に測定の順序が無関係であるということを意味している）を持っており，置換 σ に対して，同じ確率で起こるという仮定をする．この仮定の下で，α_{N+1} が全体のスコアの $100 \cdot \epsilon$ % よりも低くなることを要請する．これはつまり，

$$p(y) := \frac{\#\{i | \alpha_i \geq \alpha_{N+1}\}}{N+1}$$

が $p(y) < \epsilon$ を満たすものは棄却されるということである．

参考文献

1）K. Shin, J. Takai, J. Amari, and K. Murota. 機械学習アルゴリズム入門：類似性の科学 . 工学社 (2022).

2）S. C. Hora. *Reliability Engineering & System Safety*, 54, 217 (1996).

3）A. Der Kiureghian and O. Ditlevsen. *Structural safety*, 31, 105 (2009).

4）I. Kononenko. *Artificial Intelligence in medicine*, 23, 89 (2001).

5）T. Fischer and C. Krauss. *European journal of operational research*, 270, 654 (2018).

6）T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben. *Production & Manufacturing Re- search*, 4, 23 (2016).

7）OpenAI, :, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han,

J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O' Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. Gpt-4 technical report (2023).

8) M. Khanahmadi and K. Mølmer. *Physical Review A*, 103, 032406 (2021).

9) J. Takai, K. Shibata, N. Sekiguchi, and T. Hirano. *Physical Review A*, 107, 053308 (2023).

10) E. Hüllermeier and W. Waegeman. *Machine Learning*, 110, 457 (2021).

11) M. Kull and P. A. Flach. In *Machine Learning and Knowledge Discovery in Databases: Euro- pean Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings,Part II 14*, 18–33. Springer (2014).

12) M. Seeger. *International journal of neural systems*, 14, 69 (2004).

13) J. Denker and Y. LeCun. *Advances in neural information processing systems*, 3 (1990).

14) D. J. MacKay. *Neural computation*, 4, 448 (1992).

15) R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Busi- ness Media (2012).

16) M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. *Machine learning*, 37, 183 (1999).

17) A. Graves. *Advances in neural information processing systems*, 24 (2011).

18) Y. Gal and Z. Ghahramani. In *international conference on machine learning*, 1050–1059. PMLR (2016).

19) S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. In *Inter- national Conference on Machine Learning*, 1184–1193. PMLR (2018).

20) B. Lakshminarayanan, A. Pritzel, and C. Blundell. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. (2017).[link].

21) A. Thompson, K. Jagan, A. Sundar, R. Khatry, J. Donlevy, S. Thomas, and P. Harris. *Uncertainty evaluation for machine learning*, volume 46. NPL (2021).

22) T. M. Mitchell. *Version spaces: an ap- proach to concept learning*. Stanford Univer- sity (1979).

23) R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner- Banzhoff, and E. Hüllermeier. *Information Sciences*, 255, 16 (2014).

24) V. Vovk, A. Gammerman, and G. Shafer. In *Algorithmic Learning in a Random World*, 71–106. Springer (2022).